Why probability theory?

But it must be recognized that the notion ‘probability of a sentence’ is an entirely useless one, under any known interpretation of this term. — Chomsky (1968)

Short answer: practice proved otherwise.

Slightly long answer

- Many linguistic phenomena are better explained as tendencies, rather than fixed rules
- Probability theory captures many characteristics of (human) cognition, language is not an exception

What is probability?

- Probability is a measure of (un)certainty
- We quantify the probability of an event with a number between 0 and 1:
  - 0: the event is impossible
  - 0.5: the event is as likely to happen as it is not
  - 1: the event is certain
- The set of all possible outcomes of a trial is called sample space ($\Omega$)
- An event ($E$) is a set of outcomes

Axioms of probability state that

1. $P(E) \in \mathbb{R}, P(E) \geq 0$
2. $P(\Omega) = 1$
3. For disjoint events $E_1$ and $E_2$, $P(E_1 \cup E_2) = P(E_1) + P(E_2)$

Where do probabilities come from

Axioms of probability do not specify how to assign probabilities to events.

Two major (rival) ways of assigning probabilities to events are

- Frequentist (objective) probabilities: probability of an event is its relative frequency (in the limit)
- Bayesian (subjective) probabilities: probabilities are degrees of belief

Random variables

- A random variable is a variable whose value is subject to uncertainties
- A random variable is always a number
- Think of a random variable as mapping between the outcomes of a trial to (a vector of) real numbers (a real valued function on the sample space)
- Example outcomes of uncertain experiments
  - height or weight of a person
  - length of a word randomly chosen from a corpus
  - whether an email is spam or not
  - the first word of a book, or first word uttered by a baby
- Note: not all of these are numbers

Probability mass function

Example: probabilities for sentence length in words

<table>
<thead>
<tr>
<th>Sentence length (words)</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.155</td>
</tr>
<tr>
<td>2</td>
<td>0.185</td>
</tr>
<tr>
<td>3</td>
<td>0.210</td>
</tr>
<tr>
<td>4</td>
<td>0.194</td>
</tr>
<tr>
<td>5</td>
<td>0.102</td>
</tr>
<tr>
<td>6</td>
<td>0.066</td>
</tr>
<tr>
<td>7</td>
<td>0.039</td>
</tr>
<tr>
<td>8</td>
<td>0.023</td>
</tr>
<tr>
<td>9</td>
<td>0.012</td>
</tr>
<tr>
<td>10</td>
<td>0.005</td>
</tr>
<tr>
<td>11</td>
<td>0.004</td>
</tr>
</tbody>
</table>
Populations, distributions, samples

- In many applications, we use probability theory to make inferences about a possibly infinite population.
- A probability distribution is a way to characterize a population.
- Our inferences are often based on samples.

A sample from the distribution in the previous slide:

\[ [1, 2, 3, 3, 4, 4, 5, 7, 11] \]

Variance and standard deviation

- Variance of a random variable \( X \) is,

\[
\text{Var}(X) = \sigma^2 = \sum_{i=1}^{n} P(x_i)(x_i - \mu)^2 = E[X^2] - (E[X])^2
\]

- It is a measure of spread, divergence from the central tendency.
- The square root of variance is called the standard deviation.

\[
\sigma = \sqrt{\left( \sum_{i=1}^{n} P(x_i)x_i^2 \right) - \mu^2}
\]

- Standard deviation is in the same units as the values of the random variable.
- Variance is not linear: \( \sigma_{X+Y}^2 \neq \sigma_X^2 + \sigma_Y^2 \) (neither the \( \sigma \)).

Short divergence: Chebyshev’s inequality

For any probability distribution, and \( k > 1 \),

\[
P(|x - \mu| > k\sigma) \leq \frac{1}{k^2}
\]

This also shows why standardizing values of random variables,

\[
z = \frac{x - \mu}{\sigma}
\]

makes sense (the normalized quantity is often called the z-score).

Cumulative distribution function

<table>
<thead>
<tr>
<th>Length</th>
<th>Prob.</th>
<th>C. Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.16</td>
<td>0.16</td>
</tr>
<tr>
<td>1.0</td>
<td>0.18</td>
<td>0.34</td>
</tr>
<tr>
<td>1.5</td>
<td>0.23</td>
<td>0.55</td>
</tr>
<tr>
<td>2.0</td>
<td>0.19</td>
<td>0.74</td>
</tr>
<tr>
<td>2.5</td>
<td>0.10</td>
<td>0.85</td>
</tr>
<tr>
<td>3.0</td>
<td>0.07</td>
<td>0.91</td>
</tr>
<tr>
<td>4.0</td>
<td>0.04</td>
<td>0.95</td>
</tr>
<tr>
<td>5.0</td>
<td>0.02</td>
<td>0.97</td>
</tr>
<tr>
<td>6.0</td>
<td>0.03</td>
<td>0.99</td>
</tr>
<tr>
<td>7.0</td>
<td>0.01</td>
<td>0.99</td>
</tr>
<tr>
<td>8.0</td>
<td>0.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Variance is not linear:

It is a measure of spread, divergence from the central tendency.

Expected value

- Expected value (mean) of a random variable \( X \) is,

\[
E[X] = \mu = \sum_{i=1}^{n} x_i P(x_i) = P(x_1)x_1 + P(x_2)x_2 + \ldots + P(x_n)x_n
\]

- More generally, expected value of a function of \( X \) is

\[
E[f(X)] = \sum_{x} P(x)f(x)
\]

- Expected value is a measure of central tendency.
- Note: it is not the ‘most likely’ value.
- Expected value is linear:

\[
E[aX + bY] = aE[X] + bE[Y]
\]

Example: two distributions with different variances

Median and mode of a random variable

Median is the mid-point of a distribution. Median of a random variable is defined as the number \( m \) that satisfies

\[
P(X < m) \geq \frac{1}{2} \quad \text{and} \quad P(X \geq m) \geq \frac{1}{2}
\]

- Median of 1, 4, 5, 8, 10 is 5
- Median of 1, 4, 5, 7, 8, 10 is 6

Mode is the value that occurs most often in the data.

- Modes appear as peaks in probability mass (or density) functions.
- Mode of 1, 4, 5, 8, 10 is 4
- Modes of 1, 4, 8, 9, 9 are 4 and 9
Mode, median, mean, standard deviation

Visualization on sentence length example

Multimodal distributions

- A distribution is multimodal if it has multiple modes
- Multimodal distributions often indicate confounding variables

Another example

A probability distribution over letters

- We have a hypothetical language with 8 letters with the following probabilities

<table>
<thead>
<tr>
<th>Letter</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prob.</td>
<td>0.23</td>
<td>0.04</td>
<td>0.05</td>
<td>0.06</td>
<td>0.29</td>
<td>0.02</td>
<td>0.07</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Probability distributions (cont)

- A probability distribution is called univariate if it was defined on scalars
- Multivariate probability distributions are defined on vectors
- Probability distributions are abstract mathematical objects (functions that map events/outcomes to probabilities)
- In real life, we often deal with samples
- A probability distribution is generative device: it can generate samples
- Finding most likely probability distribution from a sample is called inference (next week)

Probability distributions

- Some random variables (approximately) follow a distribution that can be parametrized with a number of parameters
- For example, Gaussian (or normal) distribution is conventionally parametrized by its mean ($\mu$) and variance ($\sigma^2$)
- Common notation we use for indicating that a variable $X$ follows a particular distribution is
  $$X \sim \text{Normal}(\mu, \sigma^2) \quad \text{or} \quad X \sim \mathcal{N}(\mu, \sigma^2).$$
- For the rest of this lecture, we will revise some of the important probability distributions

Uniform distribution (discrete)

- A uniform distribution assigns equal probabilities to all values in range $[a, b]$, where $a$ and $b$ are the parameters of the distribution
- Probabilities of the values outside range is 0
  $$\mu = \frac{b + a}{2}$$
  $$\sigma^2 = \frac{(b - a + 1)^2 - 1}{12}$$
- There is also an analogous continuous uniform distribution

Skew

- Another important property of a probability distribution is its skew
- Symmetric distributions have no skew
- Positively skewed distributions have a long tail on the right
- Negatively skewed distributions have a long left tail
Binomial distribution

Binomial distribution is a generalization of Bernoulli distribution to \( n \) trials, the value of the random variable is the number of 'successes' in the experiment

\[
P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}
\]

\[
\mu_X = np
\]

\[
\sigma_X^2 = np(1-p)
\]

Remember that \( \binom{n}{k} = \frac{n!}{k!(n-k)!} \).

Beta distribution

Beta distribution is defined in range \([0, 1]\)

It is characterized by two parameters \( \alpha \) and \( \beta \)

\[
p(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}
\]

Where do we use it

- A common use is the random variables whose values are probabilities
- Particularly important in Bayesian methods as a conjugate prior of Bernoulli and Binomial distributions
- Dirichlet distribution generalizes Beta to \( k \)-dimensional vectors whose components are in range \([0, 1]\) and \( |x|_1 = 1 \).
- Dirichlet distribution is also used often in NLP, e.g., latent Dirichlet allocation is a well know method for topic modeling.
Example Dirichlet distributions

$$\theta = (0.8, 0.8, 0.8)$$

Gaussian (normal) distribution

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Student’s t-distribution

- T-distribution is another important distribution.
- It is similar to normal distribution, but it has heavier tails.
- It has one parameter: degree of freedom ($v$).

Joint and marginal probability

Two random variables form a joint probability distribution.

An example: consider the letter bigrams.

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>a-b</th>
<th>b-c</th>
<th>c-d</th>
<th>d-e</th>
<th>e-f</th>
<th>f-g</th>
<th>g-h</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.04</td>
<td>0.02</td>
<td>0.02</td>
<td>0.03</td>
<td>0.05</td>
<td>0.01</td>
<td>0.02</td>
<td>0.06</td>
</tr>
<tr>
<td>b</td>
<td>0.04</td>
<td>0.02</td>
<td>0.02</td>
<td>0.03</td>
<td>0.05</td>
<td>0.01</td>
<td>0.02</td>
<td>0.06</td>
</tr>
<tr>
<td>c</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>d</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>e</td>
<td>0.06</td>
<td>0.02</td>
<td>0.02</td>
<td>0.03</td>
<td>0.08</td>
<td>0.01</td>
<td>0.01</td>
<td>0.07</td>
</tr>
<tr>
<td>f</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>g</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>h</td>
<td>0.08</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.10</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Variance of joint distributions

$$\sigma^2_X = \sum_{x,y} p(x,y)(x-\mu_X)^2$$

$$\sigma^2_Y = \sum_{x,y} p(x,y)(y-\mu_Y)^2$$

$$\sigma_{XY} = \sum_{x,y} \frac{y}{y}(x-\mu_X)(y-\mu_Y)$$

- The last quantity is called covariance which indicates whether the two variables vary together or not.
- Again, using vector/matrix notation we can define the covariance matrix ($\Sigma$) as:

$$\Sigma = E[(x-\mu)(y-\mu)^T]$$
**Covariance and the covariance matrix**

\[ \Sigma = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{YX} & \sigma_Y^2 \end{bmatrix} \]

- The main diagonal of the covariance matrix contains the variances of the individual variables
- Non-diagonal entries are the covariances of the corresponding variables
- Covariance matrix is symmetric (\( \sigma_{XY} = \sigma_{YX} \))
- For a joint distribution of \( k \) variables we have a covariance matrix of size \( k \times k \)

**Correlation**

Correlation is a normalized version of covariance

\[ r = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \]

Correlation coefficient (\( r \)) takes values between -1 and 1
- Perfect positive correlation: \( r = 1 \)
- Positive correlation: \( 0 < r < 1 \)
- No correlation, variables are independent: \( r = 0 \)
- Negative correlation: \( -1 < r < 0 \)
- Perfect negative correlation: \( r = -1 \)

Note: like covariance, correlation is a symmetric measure.

**Correlation and independence**

- Statistical (in)dependence is an important concept (in ML)
- The correlation (or covariance) of independent random variables is 0
- The reverse is not true: 0 correlation does not imply independence
- Correlation measures a linear dependence (relationship) between two variables, non-linear dependences are not measured by correlation
Conditional independence

If two random variables are conditionally independent:

\[ P(X,Y|Z) = P(X|Z)P(Y|Z) \]

This is often used for simplifying the statistical models. For example in spam filtering with Naive Bayes classifier, we are interested in

\[ P(w_1, w_2, w_3|\text{spam}) = P(w_1|\text{w_2, w_3, spam})P(w_2|\text{w_3, spam})P(w_3|\text{spam}) \]

with the assumption that occurrences of words are independent of each other given we know the email is spam or not,

\[ P(w_1, w_2, w_3|\text{spam}) = P(w_1|\text{spam})P(w_2|\text{spam})P(w_3|\text{spam}) \]

Bayes’ rule

We use a test t to determine whether a patient has condition/illness c

- If a patient has c test is positive 99% of the time:
  \[ P(t|c) = 0.99 \]

- What is the probability that a patient has c given t?
  \[ P(c|t) \]

- or more correctly, can you calculate this probability?

- We need to know two more quantities. Let’s assume
  \[ P(t) = 0.00001 \text{ and } P(t|\neg c) = 0.01 \]

\[ P(c|t) = \frac{P(t|c)P(c)}{P(t)} = \frac{P(t|c)P(c)}{P(t|c)P(c) + P(t|\neg c)P(\neg c)} = 0.001 \]

Chain rule

We rewrite the relation between the joint and the conditional probability as

\[ P(X,Y) = P(X|Y)P(Y) \]

We can also write the same quantity as,

\[ P(X,Y) = P(Y|X)P(X) \]

For more than two variables, one can write

\[ P(X,Y,Z) = P(Y|X)P(X)P(Z|Y)P(X|Y)P(Z) = \ldots \]

In general, for any number of random variables, we can write

\[ P(X_1, X_2, \ldots, X_n) = P(X_1|X_2, \ldots, X_n)P(X_2, \ldots, X_n) \]

Continuous random variables

some reminders

The rules and quantities we discussed above apply to continuous random variables with some differences

- For continuous variables, \( P(X = x) = 0 \)

- We cannot talk about probability of the variable being equal to a single real number

- But we can define probabilities of ranges

- For all formulas we have seen so far, replace summation with integrals

- Probability of a range:

\[ P(a < X < b) = \int_a^b p(x)dx \]
Multivariate continuous random variables

- **Joint probability density**
  \[ p(X, Y) = p(X|Y)p(Y) = p(Y|X)p(X) \]
- **Marginal probability**
  \[ P(X) = \int_{-\infty}^{\infty} p(x, y) \, dy \]

Samples from bi-variate normal distributions

- \( \Sigma = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \)
- \( \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 0.5 \end{bmatrix} \)
- \( \Sigma = \begin{bmatrix} 0.5 & 0.7 \\ 0.7 & 2 \end{bmatrix} \)
- \( \Sigma = \begin{bmatrix} 2 & -0.7 \\ -0.7 & 0.5 \end{bmatrix} \)

Summary: some keywords

- Probability, sample space, outcome, event
- Outcome, event, sample space
- Random variables: discrete and continuous
- Probability mass function
- Probability density function
- Cumulative distribution function
- Expected value
- Variance / standard deviation
- Median and mode
- Skewness of a distribution
- Joint and marginal probabilities
- Covariance, correlation
- Conditional probability
- Bayes' rule
- Chain rule
- Some well-known probability distributions:
  - Bernoulli
  - Binomial
  - Beta
  - Dirichlet
  - Gaussian
  - Student's t

References and further reading

- MacKay (2003) covers most of the topics discussed in a way quite relevant to machine learning. The complete book is available freely online (see the link below)
- See Grinstead and Snell (2012) a more conventional introduction to probability theory. This book is also freely available
- For a influential, but not quite conventional approach, see Jaynes (2007)

References and further reading (cont.)